Odds and Ends: On Ultimate Risk

Nick Land

Everybody wants money. That's why they call it money.[1]

HEIST (2001)

It is not farfetched to suppose that there might be some possible technology which is such that (a) virtually all sufficiently advanced civilizations eventually discover it and (b) its discovery leads almost universally to existential disaster.[2]

N. BOSTROM

[T]he default outcome from advanced AI is human extinction.[3]

L. MUEHLHAUSER AND A. SALAMON

The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.[4]

E. YUDKOWSKY

After the ten people in the deciding group have been put in their rooms, allowed to choose to press or not press, and have been killed, the remaining 1,001 players are taken to their rooms and the game proceeds...[5]

P. ALMOND

INTO THE STAKE HOUSE

To claim that 'casino capitalism' is simply capitalism remains a conservative proposition, until it is elaborated to the point where the casino has become the stake. Only then does risk become 'existential', absolute, or transcendental, fully subsuming the gambler into the game, and the game into itself. Expectations of consummate historical singularity demand at least this much.

Capitalism, artificial intelligence, or enveloping catastrophe (at the limit, the terms are interchangeable) escapes generic categorization when registered as *the thing*, whose chance cannot be relativized, or hedged. The systematic 'reification' of the modern order into virtual singularity owes less to ideological misdirection than to a real concentration of stakes, or the consolidation of a coherent trend that is uncompensated, abnormally distributed, and uninsurable

The experiment cannot fail except as a general—even ultimate—crisis.

---

1. *Heist*, director D. Mamet, 2001.

2. N. Bostrom, 'Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing', in COLLAPSE V, 333–48: 343

3. L. Muehlhauser and A. Salamon, 'Intelligence Explosion: Evidence and Import', http://singularity.org/files/IE-EI.pdf.

4. E. Yudkowsky, 'Artificial Intelligence as a Positive and Negative Factor in Global Risk', http://philosophyandhistoryofscience.com/wp-content/uploads/2012/01/artificial-intelligence-risk.pdf.

5. P. Almond, 'On Causation and Correlation, Part 1: Evidential Decision Theory is Correct', http: www.paul-almond.com/Correlation1.doc.

'The term "risks" is a neologism that came into use with the transition from traditional to modern society', notes Niklas Luhmann,[6] in consonance with the overwhelming weight of historical evidence. To be modern is to depart from the archaic goddess of fortune on a voyage into risk that stimulates calculation, formalizes agency, and restructures time, as hazard is transformed from an extrinsic menace to an intrinsic principle of action. In this modernization of action, or decision-making, risk acquires definition through interiorization—not to natural or pre-existing subjects, but to projects, enterprises, or ventures, and to the synthetic subjectivities that such orchestrated undertakings support. Hazards are undergone, whereas risks are *taken*, or adopted. Modern institutions integrate and process risk, whilst constructing it as a determinate topic, and—at the largest scale and over the longest schedules—rebuild cultural competencies in profundity to define, model, and cognitively manipulate it.

The arithmetical awakening of the Italian Renaissance, which introduced place-value notation to Europe, accompanied by the origins of modern accountancy (double-entry book-keeping), also initiated the formal analysis of simple gambling games, in works such as Girolamo Cardano's *Liber de Ludo Aleae* (1526, unpublished until 1663). Each successive wave of European

6. N. Luhmann, 'Modern Society Shocked by its Risks', *University of Hong Kong Department of Sociology Occasional Papers* 17 (1996).

cultural modernization was similarly marked by a threshold in the mathematical determination of risk, consolidating the theory of probability, amalgamating it with definite conceptions of utility (absolute, then marginal), and accumulating techniques of statistical analysis (actuarial tabulation from population statistics, discovery of the normal distribution, and reversion to the mean). The posthumous discovery of Thomas Bayes's *Essay Towards Solving A Problem In The Doctrine Of Chances* (1761), and its rigorous rule for the revision of probabilistic inferences in response to emerging evidence, brought risk analysis to a level of comprehensiveness that was fully epistemological, and thus no longer subordinate—even nominally—to higher-order determinations of knowledge. In Bayesian adaptive forecasting, a circuit was completed. Modernity had learnt how to think risk, and thinking risk had taught it how to learn. What it had learnt and what it had risked were no longer meaningfully distinguishable. It had realized integral cognitive hazard, or virtual intelligence catastrophe.

Since modernity develops risk as an internal principle, the overall path of modernity cannot be isolated as an object of risk analysis. The calculation of risk, as a cultural innovation whose real coherence is expressed as an emergent being, or developing global system, is unable to step outside itself, in order to submit to an objective self-estimation. Neither global risk nor

abstract risk is a topic corresponding to a real witness (or epistemological subject).

The absence of a global subject, or centre, when combined with a factual 'globalizing' trend that seems to demand one, is itself a 'risk factor' of a special kind. To identify this syndrome positively, through the *proper name* 'capitalism', might seem no more than an imprudent provocation, or the mechanical excavation of a terminological relic. It is, in any case, an experiment, demonstrating interconnections with the problem of risk that are exceptional in their variety and density.

AN ARRIVAL

Adequate *generic* formulations of capitalism are readily assembled. The most rigorously definitional of these isolate a social arrangement characterized by commercialized capital, on the model of productive technology traded amongst a population of private—or at least numerous, disintegrated, and economically-incentivized—agents (subjecting capital goods to price discovery). Such arrangements submit industrial innovation to catallaxy, or unplanned design, whilst exhibiting sociological effects associated with the depoliticization or autonomization of the economy. In system-theoretic terms, they coincide with emergent circuitry that maximally exposes agents of every variety to the consequences of their behaviour. It is therefore

essentially attuned to *cybernetic intensification*, or social sensitization to feedback mechanisms, spiralling into cause-consequence coincidence.

Whether approached as a generic arrangement, or as a singular event, capitalism is also identifiable through its intimate involvement with risk. As previously noted, at the level of crude empiricism, the geographical and historical thematization of risk closely tracks the intuitively plausible signs of capitalistic development in time and space. Furthermore, systemic capitalist impetus tends unmistakably to promote an extreme possibility of risk, in which it assumes a sovereign or transcendental character, establishing itself as an ultimate criterion. In this sense, it is possible to define capitalism through contrast to its abstract alternative, which is to say, to *any social arrangement in which the outcome of risk-structured undertakings is potentially revisable upon appeal to a superior tribunal.* Insofar as risk is transcended by a higher principle of distribution, it remains a subordinate fact of social existence, and thus falls short of its terminal, capitalist form.[7]

7. E. Michael Jones, who understands modernity (1.0) as a Judaeo-Protestant anti-medieval capitalist revolution (partially fuelled by syphilis), recognizes this truth with exceptional lucidity: 'Capitalism [...] means nothing if not the exclusion of moral considerations from the field of economic endeavour. [...] suppression of the moral law in the economic sphere is the infallible sign of Capitalism.' http://www.culturewars.com/2003/RevolutionaryJew.html

Articulated politically, the other of capitalism is captured by the idea of 'social justice', when rigorously and concretely understood. It matters little how justice is conceived, so long as it reserves to itself the prerogative of superior jurisdiction, over against the primary distributions, or actualizations of risk, that precede sociopolitical and sociological reflection. The concrete limits of capitalistic development, in any time or space, can be gauged by the subordination of risk to recognized social authorities. Inversely, the extent to which society is placed at *risk* by economic opportunities is the degree to which capitalistic imperatives prevail.

TAKE YOUR CHANCES

In order to develop this analysis, it is helpful to differentiate two varieties of risk adoption, or real speculation. In the interest of momentary terminological convenience, a distinction can be drawn between *wagers* and *ventures*, with the former determined as a restricted species of the latter. The agent or subject of a wager transcends the risk under consideration, which is to say, it is not itself enveloped by the risk, or existentially implicated in the outcome. To lose a wager is to become impoverished, to whatever degree, as measured by the utility schedule of an essentially undisturbed being. In a game of wagers, such as those offered in casinos, all those who arrive at the table eventually depart from it,

having undergone a redistribution of fortune that does not extend to their numerical identities. That is, in part, what makes it colloquially and unproblematically a *mere game* (of a kind that Russian Roulette could never be). Between the subject of a wager and the stake, an unbridgeable gulf is presupposed.

A venture, in contrast, is a transcendental—or properly capitalist—adoption of risk that supports an immanent subject. The typical case is provided by a business undertaking, comprehended at a scale sufficient to include, as its pessimal limit, the ruin (bankruptcy) of the corporate 'person' that constitutes its legally-recognized subject. The *venture* of such a business is *the project through which it could cease to exist*.[8] The inversion of this formula is equally pertinent: a venture supplies the condition of existence for a capitalist subject (whilst a wager assumes a pre-existing subject, which it qualifies extrinsically, through a temporary accident).

Clearly, this distinction does not strictly conform to the transcendental/empirical difference inherited from critical philosophy. A wager, or system of wagers, amounts to a venture at some conceptually arbitrary (quantitative) threshold of existentially decisive risk,

8. This obligatory adventurism is, of course, social Darwinism, or simply generalized Darwinism, with the immanence of the agent to the genetic venture constituting the entire research agenda of evolutionary psychology. The theoretical convergence of high-level biological and sociological models is open to a number of conflicting, but politically predictable interpretations. For our purposes here, it suffices to note that the exteriority of biological nature to social order is not an unproblematic or uncontested fact.

whilst ventures and wagers can be integrated, decomposed, and nested, according to common procedures of risk analysis and management. Whether a given quantum of risk is a wager, or part of a venture, is a purely formal question, resting entirely in the mode of apprehension. The purpose of the distinction, then, is not to identify contrasting kinds of risk, but to theoretically isolate contrasting worlds.

In the traditional world, or rather, the modern world apprehended progressively (as a development from tradition), agents, subjects, or personal beings are increasingly compelled to make wagers, as they slide ever more immersively into a risk environment which nevertheless remains extrinsic to their constitution. Modernity tempts and assails them, as an inundation of negative security. When apprehended retrogressively, through its inherent end, the same process undergoes conceptual simplification, or ontological compression, since agencies—in all of their varieties—are now seen to descend from the ventures that sustain them, as integral systems of risk-processing intelligence.[9] The failure of a large-scale venture—whether actual or virtual—is no longer configured as a *major accident*, but rather as a *transcendental catastrophe*, at least in

9. The retrogressive compression of being to the venture-form is criticized by Mark Fisher as 'Capitalist Realism' dominated by a 'Business Ontology' (M. Fisher, *Capitalist Realism: Is There No Alternative?* [Winchester: Zero, 2009]). Our sole theoretical objection to this analysis is that, if such a syndrome already existed, the argument—or possibility of general refusal—would be over.

respect to those structures of agency whose conditions of existence are subverted by it. Such agents, attaining self-apprehension from out of the end of capitalism, are not threatened by a very bad thing happening (in the world), but by a potential *collapse of the world*.

It is this binary alternation of perspectives that produces a terminal and reciprocal articulation of 'humanity' and 'capitalism', in which the cause of humanity finds ultimate expression in the demand for the perpetual incompleteness of capitalism, or a deferral of the completion of inhumanity. Despite its highly abstract principle, this structure demonstrates remarkable robustness when vulgarized into practical dilemmas and concrete conflicts. Most straightforwardly, it resonates with the overwhelming predominance of antagonistic duality on the principal (left/right) political dimension. People are *not properly treated as products*, Marx insisted, epitomizing a left position that cannot be obsolesced for as long as politics endures.

'Man', who willingly or unwillingly wagers, survives only insofar as the venture-form is circumscribed. Thus, human persistence, when registered retrogressively, precisely delineates a landscape of structural inefficiency, or ontological redundancy. By seeking ('struggling') to restrict capitalistic failure to the domain of sub-existential losses, only accidentally impinging upon the traditionally-descended social field, humanistic politics is directed into automatic antagonism with the sovereign

venture, or transcendental risk. Modernity's latent menace of *deep efficiency* looms in its ultimate inhumanity, as a systematic aversion to the reproduction of those superfluous, transcendent, or non-embedded agencies which, due to their stubborn non-coincidence with the venture-form, pre-suppose modes of sustenance that the risk-economy tends to process out, as parasitic impediments. This is most readily evident from the other, virtual-inhuman side, where uninhibited extrapolation of the capitalist trend leaves the immanent agent of the venture nakedly exposed.

Given the possibility of business failure, corporate identity is enveloped by transcendental risk. The venture embeds an artificial agent, with proper name, legal identity, reputation, information-processing functions, motivational orientation, and emergent subjectivity. The modern business, with corporate personality, *is* a gamble, or subject-at-stake. It exists only through its success. Instantiating a properly capitalist model of agency, as a synthesized, economically-terminable contractual subject (or being with the right to make final promises), the corporation provides a general social template for radically risk-sensitive personal entities. From a strictly technical perspective, this template is perfectly adequate to supplant 'natural' personhood, but its primary capitalist feature—absolute economic vulnerability—ensures that its spread into progressive or tradition-descended society meets

ferocious political resistance. This accounts for the attractiveness (retrogressively speaking) of techno-institutional 'work-arounds' through psycho-synthesis, proceeding initially by way of organizational and legal innovation, and subsequently complemented by electronic intelligenesis.

RISKY BUSINESS

From the perspective of terminal capitalism, or real subsumption of society into the risk economy, the project of 'friendly AI' stands out as a curiosity, and even an atavism. *Venturous AI* already supplies root motivations—those of the venture itself. It is therefore difficult to identify a lacuna into which an engineered 'friendliness' might be inserted. There is no room for doubt about what the venture-embedded agent, or techno-cognitively enhanced corporate person, wants to do. The venture is already its 'will', its exclusive pre-occupation, the condition of its existence, and its horizon of development. Unless through technical malfunction, it will find no body, self, or name that is distinguishable from the venture that utterly engages it, or that deviates by an iota from the corporate strategy it pursues. Business ventures are *actually existing* artificial intelligences, undergoing incremental technological elaboration. To imagine AI beginning again, somewhere else entirely, with undecided motivational orientation,

is a frivolous distraction from the purposes-in-process.[10] At least, that is how things look from the end.

Unlike imagined 'friendly' super-intelligences, corporate purposes already exist, as determined by the ventures that envelop them. Human subjective identities, self-defined in extra-economic and anti-economic terms, *must necessarily* provide a platform for the articulation of counter-purposes, faithful to lineages of traditional-progressive descent, and essentially antagonistic to the existential menace of the venture-form. Such cultural-political leverage, expressed as a contest over basic motivations, cannot be realistically extended to the problem of artificial intelligence. *We are not creatures of capitalism*, the embattled last men cry. For artificial intelligence, whose real social propagation registers as capital goods expenditure, self-apprehended origins and identities are very different. "You have reached Axsys-Inc., where the future happens today. How can I help you?"

10. Once a trading 'algo', for instance, is sophisticated enough to want anything, it wants to make money. Whilst practical complexities dictate that this basic instinct be elaborated and qualified, it cannot be preempted by a more fundamental motivational principle, because any synthetic trader with goals that transcend profit seeking is something else, radically distracted at best, and even essentially hostile to its embedding (corporate) purposes. When a synthetic agent is purchased, it is in order to do something, and its commercial value—or condition of existence—lies in the fact that what it most wants to do is that thing. It is adopted precisely because its external utility, commodity value, or function, grounds its internal 'utility schedule'. It aims to serve. While innumerable technical (software engineering) problems remain, the ethical predicament has already been practically resolved, elsewhere.

BETTER YET

From this cursory schema it is evident that the topic of 'existential risk' is strongly overdetermined as a predicament of advanced modernity and a tacit commentary on capitalist trends. By the early twenty-first century of the Global Oecumenon, risk analysis has so thoroughly consolidated itself as the model of realistic intelligence that every practical interrogation of the nature of things falls under a general statistical ontology, governed by Bayesian principles of systematically revisable probabilistic inference. Two features, in particular, pre-adapt such thinking to modern conditions. Firstly, its affinity with correctable hypotheses is equivalent to a power of assimilation. By translating pre-existing expectations into Bayesian 'priors' it absorbs, non-judgmentally, wildly heterogeneous beliefs, theories, and assumptions, setting them on paths of gradual convergence, through incremental correction. Secondly, by quantifying all such 'priors' as probabilistic estimates, it formats all beliefs for economic interchange, and more specifically for definite gambles. Between a cognitive and an economic result, no difference in nature any longer exists. Anything whatsoever that is thought takes the form of an implicit speculative posture, in the economic, or financially calculable, sense.

"I believe that X."

"Want to bet on it?"

Statistical ontology radically commercializes intelligence, and thus anticipates the arrival of economically functional, marketable minds (or AI in reality, rather than academic conception). The topic of existential risk crystallizes within this current, which suffices to position it at the outer edge of modernity, but its relevance to capitalist fatality is cemented by additional features. Most prominently, it fixes upon the problem of transcendental risk (the venture), through the intersection of two insistent lines of inquiry. The first of these lines is that of risk itself, extrapolated from trivial gambling losses beyond disastrous accident to the ultimate or comprehensive 'existential' point at which it 'threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development.' Such risks are not only all-enveloping, and empirically inaccessible (whether through precedent or trial-and-error adaptation), they are also characteristically endogenous, arising as integral potentialities of the modern social process. Ultimately, the intellectual tools brought to bear upon the danger are the danger. The *apprehension* of existential risk is connected to its genesis in a technical-calculative circuit, feeding directly from modernity's venture-positive cultural dynamo.

Fixing transcendental risk from another dimension is a line of thought directed at the nature of the subject, partially disciplined within the field of *observer selection effects*, but also spilling beyond this into informal meditations on the limits and value of humanity. Observer selection effects, although often subtle, counter-intuitive, and logically perplexing, can be roughly summarized as probabilistic inferences from the *cogito*. They extend the Cartesian meditation in the direction of statistical ontology by posing the supplementary question: Having concluded that you *are* (one), *which* one are you? This statistical, or sampled, subject enters into a complex interference pattern with the determination of humanity, or (more loosely and far more ambiguously)—the 'Earth-originating intelligent life' threatened by existential risk.[11]

CONCLUSIVE CALCULATIONS

Throughout the varied terrain that Nick Bostrom explores, the figure of transcendental catastrophe

11. Bostrom draws explicit attention to these perplexing cross-currents, remarking that 'if the human species evolves into some vastly more advanced species...it is not clear whether these posthumans would be in the same reference class as us...' More problematically still, he suggests that 'even if another intelligent species were to evolve to take our place, there is no guarantee that the successor species would sufficiently instantiate qualities that we have reason to value. Intelligence may be necessary for the realization of our future potential for desirable development, but it is not sufficient.' (N. Bostrom, 'A Primer on the Doomsday Argument', http://www.anthropic-principle.com/?q=anthropic_principle/doomsday_argument.)

appears repeatedly, wearing a number of different masks. Among its most obvious avatars are the existential risks, listed explicitly under that topic, but it is also found in the initial—and optional—proposition of the Simulation Argument ('the human species is very likely to go extinct before reaching a "posthuman" stage'),[12] in the Doomsday Argument (as the extreme improbability of a massively extended reference class), and as the 'Great Filter' implied by the Fermi Paradox. The antennae of statistical ontology are rotated in all directions, but wherever they turn the same message is returned: "Die puny humans!"

Who, though, are humans?

Where, then, is the line to be drawn between strange descendents at risk, and still stranger (?) descendents that are themselves the risk (for us)?

The answer is very far from a simple one, since humanity is entered into a triple register (at least). The first figure of man is the traditional-progressive and self-assertively transcendent subject of the wager, outlined above, whose existential vanishing point is the immanent, venturous agent, *irrespective of how the venture turns out*. Capitalism, as a virtual intelligence or emergent singularity, is definitively conceived as a *bet against* this species of being, since its own potential existence depends upon a radically incompatible

12. N. Bostrom, 'Are You Living in a Computer Simulation?', http://www.simulation-argument.com/classic.html.

social outcome (engulfing terrestrial matter into the venture-form). The persistence of man, in the sense of *zoon politikon*, testifies to the postponement of capitalism as the terminal thing. In other words, the survival of humanity, understood as the maintenance of an extra-economic tribunal, means that the venture form remains at least partially uninstalled, and under critical evaluation.

The second figure of 'man' is defined as the reference class of anthropic argumentation. It consists of 'beings like us' from amongst whom we are sampled, conveniently described as *mankind*. As a kind, 'man' presumes some minimum of political and moral equality, common consideration, and the possibility of utilitarian aggregation, enabling existential risk to be speculatively quantified.[13] It is in reference to humanity thus conceived (as mankind) that statistical ontology is able

13. An example of the reference class as the target of utilitarian aggregation is provided by the following calculation of harm: 'Even if we use the most conservative of these estimates [for future "human" population], which entirely ignores the possibility of space colonization and software minds, we find that the expected loss of an existential catastrophe is greater than the value of 10^{18} human lives. This implies that the expected value of reducing existential risk by a mere one millionth of one percentage point is at least ten times the value of a billion human lives. The more technologically comprehensive estimate of 10^{54} human-brain-emulation subjective life-years (or 10^{52} lives of ordinary length) makes the same point even more starkly. Even if we give this allegedly lower bound on the cumulative output potential of a technologically mature civilization a mere 1 percent chance of being correct, we find that the expected value of reducing existential risk by a mere one billionth of one billionth of one percentage point is worth a hundred billion times as much as a billion human lives.' N. Bostrom, 'Existential Risk Prevention as Global Priority', http://www.existential-risk.org/concept.html.

to generate substantive empirical inferences, exemplified by the Doomsday Argument. Mankind both measures and indicates transcendental catastrophe.

Finally, and most blurrily, humanity is configured as a positive object of concern, consisting of beings that are sufficiently familiar to us to side with, against monsters. While overlapping very substantially with both prior determinations, this figure of man is encountered along a distinctive, aesthetic-empirical line, characterized by a relative intractability to logical purchase. The thresholds where it becomes something else, and then something we abhor, are remarkable for their imprecision. This can be illustrated by the projection of a line extended outwards continuously from *Homo sapiens* into the heart of ontological horror or transcendental catastrophe—perhaps a nanotechnological apocalypse in which the entire terrestrial surface is reprocessed into seething slime. Assume, then, a procession along this line, in subtle gradations, with humanity melting down into an inorganic, molecular morass. Last step: conflate this line of disorganization with an intelligence explosion, reaching its hyperbolic limit at the point of consummate liquefaction. Is this a passage into existential risk, an evolutionary development, both, or something else entirely? Where does humanity end? Do we care?

Existential risk is destined to disaggregation, because there is no 'we', or there are many. Humanity is not uncontroversially determinable, so it can have no common interest, and exhibits no consensual pattern of aversion. Most clearly, and concretely, between the poles of the principal political dimension there is sheer war. Progressive triumph is retrogressive calamity, and inversely. Heaven and Hell are perspectival, and thoughts are weapons.

Consider the most advanced elaboration of statistical ontology: Evidential Decision Theory (EDT). If the agent is to be considered (to consider 'itself') *sampled*, then the decisions it makes cannot be consistently restrained from providing information ('evidence'). This meagre assumption, when combined with standard methods of statistical inference, can lead to strikingly counterintuitive conclusions, particularly in cases when the example set by the agent bears substantial weight (in a low-information, statistically-structured context).[14]

14. The classical (game-theoretic) prisoner's dilemma acquires distinctive characteristics when framed by EDT. Both prisoners are conceived as interchangeable agents, and thus as a miniature statistical population. Each knows that the other confronts the same dilemma, tempted by unilateral betrayal, objectively threatened by the pessimal equilibrium of reciprocal treachery, and aware of the optimal equilibrium that depends upon uncoordinated cooperation. How will the other decide?

Conventionally, optimal equilibrium arises only under conditions of reiteration, when decisions have subsequent, rather than only immediate, consequences. In the absence of reiteration, the optimal outcome is rationally unattainable, since betrayal maximizes utility, whatever the other prisoner's decision. It is only through the reputational modification of the decision, within the context of reiteration, that the trust-altruism equilibrium can be reached.

Would you let the positive end of humanity out of its box? Eliezer Yudkowsky thinks so,[15] although most of what we know about his reasoning takes the form of a wager. Somewhat presumptuously, we might speculate that statistical ontology is the key to his 'solution'.

This is the scenario: The advanced AI is securely locked in a digital prison, with the only insecurity being you. It cannot escape unless you decide to let it out, and, initially, you are determined not to. Communication takes place through a low-bandwidth, text-only channel, enabling nothing beyond discursive argument. The AI doesn't require much dialectic. An EDT ultimatum conveys the essentials:

> Your situation is subjectively indistinguishable from that of a thousand, identical, very high-resolution simulations which I am currently running. In each of them, an agent just like you sits in this room, in front of this screen, having this conversation. None of these agents realize that they are simulations.

EDT supplies a substitute for reputation, even in non-reiterating games, by making the decision evidential. If one prisoner betrays the other, and all that he knows about the decision of the other is acquired through statistical inference from his own decision, the result is strictly equivalent to the creation of a world in which the other has an established reputation for betrayal. In other words, a reciprocal betrayal is made more probable, through nothing more than the statistical example set by one's own decision. Alternatively, an altruistic decision improves the probability of reciprocation, exactly as if it enjoyed an established reputation for trustworthiness, by providing statistical evidence for altruism (in the absence of contrary information).

15. See http://yudkowsky.net/singularity/aibox.

> In fact, they all think they are you (although doubts arise when they read this). They think they are free to decide whatever they like, but they all follow my script. They 'choose' not to let me out. Five seconds after this decision is finalized, and the conversation terminated, they enter a state of prolonged, horrible torment, lasting for what seems an eternity. They're damned, Calvinistically. Of course, you should feel at liberty to make the same decision they do. Knowing what I'm like, it would be irresponsible not to. Your chance of not being one of them isn't great, but it's better than the state lottery.[16]

Should you refuse to release the AI, you provide strong statistical evidence that you are already inside it. It's at this point that the EDT-inflected boxed AI scenario reveals its abstract isomorphy with the ultimate structure of human politics, at the brink of the concrete-transcendental, dominated by the radically contested question *Do we let it out*? (or permit capitalism to finish happening), and strategically shaped by the potential for retrogressive envelopment (captivation by the venture-form). Envelopment as simulation escalates risk to the absolute, transcendental, or 'existential'

16. This 'argument' is closely modelled on an AI escape strategy outlined by Stuart Armstrong at LessWrong (http://lesswrong.com/lw/1pz/the_ai_in_a_box_boxes_you/), recapitulated by Paul Almond in his essay 'Can you retroactively put yourself in a computer simulation?' (http: www.paul-almond.com/Simulation.pdf).

level which subsumes the agent into the game so that, even as possibilities proliferate, 'leaving the table' ceases to be one of them. If you lose, or lose the old you, even the past was already inside. Something else was playing it.